多专长专家识别方法研究*
——以大数据领域为例

■ 刘晓豫 朱东华 汪雪锋 黄颖
北京理工大学管理与经济学院 北京 100081

摘要：[**目的/意义**]国家政府、大中型企业以及研究机构面对技术难题，如何找到合适的专家是迫切需要解决的问题。面对需要运用多学科知识来解决的综合性复杂难题，寻找到多专长专家显得尤为重要，寻找合适的方法识别出多专长专家是本研究的目的。[**方法/过程**]利用专家所发表的学术论文数据，通过抽取专家有代表性的研究专长特征，基于TFIDF加权的重叠K-means聚类算法对专家进行重叠聚类划分，挖掘出专家的多个研究专长，进而识别出多专长专家。[**结果/结论**]研究结果表明TFIDF加权的重叠K-means聚类算法在查准率、召回率和F值上有良好的表现，可以识别多专长专家。

关键词：专家识别　重叠K-means　多专长专家　大数据　TFIDF
分类号：G316
DOI：10.13266/j.issn.0252-3116.2018.03.007

1 引言

在日益激烈的国际化与商业化竞争中，迅速地掌握与分析需求并提供高效的解决方案是取得决胜的关键因素。在当今知识社会，迫切的知识需求正在显现，专家识别与推荐作为信息检索和知识管理领域的研究热点，已经越来越受到人们的关注。专家识别目的是通过一系列的方法来发现那些拥有丰富专业知识、技能与经验的领域专家，以便组织团队，指导研发，攻关技术，以此来提高工作或生产效率[1]。面对目前国家政府、大中型企业以及研究机构技术专家无处选、无法选的现状，如何针对特定知识与技术需求找到合适的专家是一个值得研究和急需解决的问题。

专家信息的分析和挖掘是专家识别的前提，如何表征专家知识，对专家进行聚类和分类是其中关键的一环。专家专长是指专家对技能与知识的掌握[2]，R. Glaser[3]指出专家专长具有5个特征：具有持续竞争力；具体；可以更好地识别重要且有意义的模式；有一定程序以及便于通过事实识别问题。专家专长是一种隐性知识，故而要通过有形的结果，如专家主持项目、发表期刊论文等相关文档进行专家专长的识别和挖掘[4]。以往的研究常以专家最高产的研究领域代表其研究专长，但是在实际情况中，专家往往具有多个研究专长[5]。发现和识别出专家的多个研究专长能更好地进行专家评估和推荐。此外，以往的研究大多使用非重叠聚类方法对专家进行聚类，这种方法把专家唯一分配到某一类别中，忽略了专家的多个研究专长，不能识别多专长专家。针对现有研究中存在的这一问题，本文采用重叠聚类算法对专家进行聚类，避免非重叠聚类带来的信息缺失问题，同时更好地表征专家专长，挖掘多专长专家。

鉴于此，本文采用专家发表的学术论文为数据，通过向量空间模型来表征专家知识，利用TFIDF（Term Frequency - Inverse Document Frequency）加权的重叠K-means聚类算法[6]对专家进行聚类，同时识别多专长专家。本文以大数据领域的专家为案例进行研究，识别出大数据领域的多专长专家。

2 文献综述

专家专长的识别是专家遴选和推荐的基础，早期的专家专长识别方法多依赖于专家本人描述自己的专长领域，并以此构建数据库，再利用传统数据库查询语

* 本文系国家自然科学基金面上项目“开放数据环境下技术专家定位与评估方法研究”（项目编号：71673024）研究成果之一。
作者简介：刘晓豫（ORCID：0000-0003-2509-8457），博士研究生，E-mail：xiaoyu.liu2019@foxmail.com；朱东华，教授，博士生导师；汪雪锋，教授，博士生导师；黄颖，博士研究生。
收稿日期：2017-08-26　修回日期：2017-11-20　本文起止页码：55-63　本文责任编辑：王善军

言来识别专家专长,该方法的主要缺点是专家参与的主观性以及对数据库的更新缺乏时效性[7]。对此,一些学者尝试使用文档数据(论文、专利、项目报告等)[8-10]及其他行为数据(网络标签、社交小组等)[11]来分析专家专长。从方法上看,目前的专家专长识别研究主要通过基于本体的方法、基于拓扑结构社区发现算法和基于主题的专家聚类方法来实现[12]。

基于本体的专家专长识别方法通过构建领域本体,可以很好地解决关键词之间的语义关系[13-14],从而实现专家专长识别。胡月红等[15]利用 FCA 和关联规则分析的方法构建了情报学领域的本体,利用关键词到本体之间的映射实现了基于关键词的专长描述到基于领域本体的专家专长描述;刘昕民等[16]提出了 4 层模糊本体扩展框架,并利用该模型建立了科技评价领域的专家模糊本体,实现了专家遴选。

基于拓扑结构的方法从网络模型的拓扑结构出发,将专家视为网络节点,将专家之间的联系当作网络的边,从而建立起网络模型,如作者合著网络、作者耦合网络[17]、作者共引网络等;Y. Li 等[18]利用香农熵计算网络信息,通过引用网络进行专家社群挖掘;B. Dom 等[11]利用基于图论的排名算法进行专家社群分析;巩军等[19]利用谱分割算法和模块度评价指标对专家专长进行划分;刘萍等[20]利用共词网络,以社区划分的方法对关键词聚类,识别专家专长。

基于主题的专家聚类[21-22]方法通过文本挖掘,可以找到专家的研究兴趣和范围,挖掘兴趣相投的专家聚簇[23],主要的算法包括识别研究主题的层次结构的层次聚类[24]、LDA[25]、PLSA[26]等主题模型。张晓娟等[7]利用 PLSA 识别图书情报领域的专家专长,通过文档-主题和主题-关键词两个矩阵,确定专家的研究主题。

通过本体进行专家专长识别往往要耗费大量时间和精力来构造领域本体,通过拓扑结构发现专家社区往往不能很好地表示出社区和社区之间的关系,专家专长识别也缺乏对专家研究内容的分析。通过主题进行专家聚类和识别不仅能更好地表达语义,而且在处理大量数据时有明显的优势。故而,本文选用基于主题的专家聚类方法进行专家专长识别。

对于基于主题的专家聚类方法,如何通过文本表征专家研究专长,以及如何计算专家对不同主题的隶属度是影响专家专长识别效果的两个重要因素。专家发表论文的关键词、项目的申请书以及网络标签等都被用来表征专家研究专长,其中论文的关键词可以具体、准确地表征科研领域的专家专长,故而被广泛使用。在对专家进行聚类的过程中,传统的聚类算法往往把专家分入唯一的类别当中,而实际情况中大多数专家都具有多个研究专长,传统的非重叠聚类忽略了这一情况。因此,如何避免聚类过程中造成的信息损失是专家专长识别研究中的一个重要问题,本文考虑引入重叠聚类的思想解决这一问题。重叠聚类算法[27-28]的主要思想是挖掘每个对象对不同类别的隶属度,通过设置合适的阈值更好地对每个对象的特征加以表示,使分类结果更准确全面,同时具备更强的可读性。基于主题的专家聚类方法也由此衍生出基于主题的专家重叠聚类方法。当前基于主题的专家重叠聚类方法的研究相对缺乏,本文试图在这方面做出尝试。

3 研究方法

通过基于主题的专家聚类方法对专家专长进行识别,首先需要使用恰当的主题词表征专家知识,这是专家专长识别的基础;其次需要合适的聚类算法来计算专家对于每个主题的隶属度,这是专家专长识别的关键。本部分从专家-关键词矩阵构建和重叠聚类算法分析两个部分对于研究方法进行阐述。

3.1 专家-关键词矩阵的构建

专家发表的论文、专利、项目等文本信息包含着丰富的专家知识,如何从中获取有效的信息,提升知识发现的能力是当前情报学研究中十分关注的问题。通过对专家文本进行筛选和处理,挑选出合适的主题词表征专家知识,然后通过一定规则构建专家-关键词矩阵是专家专长识别的第一步。

3.1.1 关键词的获取 情报学、科学学等领域中大量研究者对文本数据的处理展开了研究,大量的分析技术和方法均围绕主题词展开[29]。首先,通过科技文献获取主题词,这些主题词可以是专家自己提供的关键词,也可以是通过自然语言处理得到的关键词。其次,使用停词表、科技期刊文献常用词表等移除主题词中无意义或文献中普遍出现的单词。然后,基于单词的词干通过模糊语义处理对主题词的动名词变化、单复数变化以及时态变化等做清洗,同时通过人工建立缩写词表,将主题词中的全称与缩写进行合并。最后,依据主题词中出现频次或者 TFIDF 等方法选择合适的关键词表示领域的专家知识。

3.1.2 专家-关键词矩阵的构建 基于获取的表征专家知识的关键词,可以构建出专家-关键词的共现矩阵[30-32]。通过对特征项赋权的方法增加特征项的

区分能力是文本分类中一种常用的方法,有研究显示文本分类中特征项权重的赋予对于分类效果有较大的影响[33]。在本文的研究中,考虑到不同的关键词在进行专家聚类的过程中的区分能力不同,对其赋予不同的权重以期得到更好的聚类效果。TFIDF 是一种用于信息检索与数据挖掘的常用加权技术[34],其主要思想是:如果某个词或短语在一篇文章中出现的频率高,并且在其他文章中很少出现,则认为此词或者短语具有很好的类别区分能力,适合用来分类。实际上,TFIDF = TF * IDF,其中 TF 为词频(Term Frequency),IDF 为逆向文件频率(Inverse Document Frequency)。

本文以专家为行向量,关键词为列向量构建如下的专家-关键词矩阵:

$$M = \begin{matrix} tf_{11} & \cdots & tf_{1p} \\ \vdots & \ddots & \vdots \\ tf_{n1} & \cdots & tf_{np} \end{matrix},$$

其中 tf_{ij} 表示专家 i 所发表的文献中主题词 j 出现的频率。对关键词计算 TFIDF,得到向量 $TFIDF = tfidf_1$, $tfidf_2$, $\cdots$, $tfidf_p$),其中 $tfidf_j$ 表示主题词 j 的 TFIDF 值。最终,基于 TFIDF 加权的专家-关键词矩阵可以表示如下:

$$M' = \begin{matrix} tf_{11} * tfidf_1 & \cdots & tf_{1p} * tfidf_p \\ \vdots & \ddots & \vdots \\ tf_{n1} * tfidf_1 & \cdots & tf_{np} * tfid_p \end{matrix}。$$

3.2 重叠 K-means 聚类算法

本文选取了 G. Cleuziou[6] 提出的重叠 K-means 算法对专家进行聚类划分[35],并对算法进行加权改进。区别于传统的 K-means 算法,重叠 K-means 算法将每个数据点聚类到一个或多个聚类当中。重叠 K-means 算法的优点在于:①相对于分配聚类,重叠聚类算法可以把一个点分配给多个聚类;②可以更客观地反映点的位置,因为算法的停止条件是每个点的影像离这个点的距离足够小;③数据处理相对连续,在图像识别领域有广泛应用;④算法复杂度低,对大数据量的数据有时间优势。基于这些优势,本文使用重叠 K-means 聚类算法对专家进行聚类。这种方法可以弥补以往的专家专长识别研究中往往只识别专家一个专长的不足,避免非重叠聚类带来的信息缺失问题,全面挖掘专家的研究专长,更好地识别多专长专家。

重叠 K-means 聚类算法分为两个过程,聚类过程和点的分配过程。

3.2.1 聚类过程　聚类过程通过迭代不断更新聚类中心,并达到组内差异最小化、组间差异最大化。每位专家用 p 维向量 $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ 表示,专家集记为 $X = \{x_i\}_{i=1}^{n}$,将 n 位专家聚为 k 个重叠聚类具体步骤如下。

①随机选择 k 个初始点作为 k 个聚簇的初始中心点,记为 $\{m_c^{(0)}\}_{c=1}^{k}$。

②计算每个 x_i 到 k 个初始点的距离,将其分配到距离最小的这一组中,得到一个 X 的覆盖 $\{\pi_c^{(0)}\}_{c=1}^{k}$,其中 $\pi_c^{(0)} = \{x_i | m_c^{(0)} \in A_i^{(0)}\}$, $A_i = \{m_c | x_i \in \pi_c\}$ 表示 x_i 属于的聚簇的集合。

③令 t = 0。

④对每个聚簇 $\{\pi_c^{(t)}\}$,计算新的各个聚簇中心点 $m_h^{(t+1)} = \mathrm{PROTOTYPE}(\pi_c^{(t)})$。

⑤进行新的聚类分配,计算分配函数 $A_i^{(t+1)} = \mathrm{ASSIGN}(x_i, \{m_c^{(t+1)}\}_{c=1}^{k}, A_i^{(t)})$,得到新的覆盖 $\{\pi_c^{(t+1)}\}_{c=1}^{k}$。

⑥计算目标函数值 $I(\{\pi_{t+1}\}) = \sum_{x_i \in X} \mathrm{dist}(x_i, \varphi(x_i))$,如果 $t_{max} > t$ 或者 $I(\{\pi_t\}) - I(\{\pi_{t+1}\}) > \in$,令 t = t + 1,返回 step4;否则结束循环,输出 $\{\pi_c^{(t+1)}\}_{c=1}^{k}$。

3.2.2 点的分配过程　点的分配过程,也就是分配函数的计算过程,通过每位专家在每次迭代中对各个类别的隶属度,找到最合适的分配方案。

①记 $A_i = \{m^*\}$,其中 $m^* = \underset{\{m_c\}_{c=1}^{k}}{\mathrm{argmin}}(\mathrm{dist}(x_i, m_c))$,计算点 x_i 的 $\varphi(x_i) = \frac{\sum_{A_i} m_c}{|A_i|}$。

②寻找除了已经分配到的聚类之外最近的中心点,$m' = \underset{\{m_c\}_{c=1}^{k} \backslash A_i}{\mathrm{argmin}}(\mathrm{dist}(x_i, m_c))$,计算在新分配 $A_i \cup \{m'\}$ 下点 x_i 的 $\varphi(x_i)$。

③如果 $\| x_i - \varphi'(x_i) \| < \| x_i - \varphi(x_i) \|$,令 $A_i \leftarrow \{m'\}$, $\varphi(x_i) = \varphi'(x_i)$,返回②;否则,计算原始分配 A_i^{old} 下的 $\varphi^{old}(x_i)$,如果 $\mathrm{dist}(x_i, \varphi(x_i)) < \mathrm{dist}(x_i, \varphi^{old}(x_i))$ 则输出 A_i,否则输出 A_i^{old}。

其中,计算中心点的方法如下:$m_h^{(t+1)} = \mathrm{PROTOTYPE}(\pi_c^{(t)})$, $m_h^{(t+1)} = \frac{\sum_{x_i \in \pi_h} \alpha_i m_h^t}{\sum_{x_i \in \pi_h} \alpha_i}$,其中 $\alpha_i = \frac{1}{|A_i|^2}$, $A_i = \{m_c | x_i \in \pi_c\}$ 表示 x_i 属于的聚簇集合,$|A_i|$ 表示 x_i 属于的聚簇的集合个数,数据点 x_i 在 h 聚簇中的映射 $m_h^i = |A_i| x_i - \sum_{m_c \in A_i \backslash \{m_h\}} m_c$。

通过重叠 K-means 聚类算法对专家进行聚类,可以得到每个专家所属的类别。每个类别表示一个研究专长,每个专家所属的类别即为其具有的研究专长,属于多个类别的专家即为多专长专家。

4 案例分析:大数据领域的多专长专家识别研究

4.1 实证数据选取

本文选取 Web of Science 核心集中被 SCIE/SSCI 索引的大数据领域的相关论文作为研究对象。检索策略的好坏关系到检索结果的质量,从而影响到最终分析的准确程度。因此,在查阅大量文献后,本文借鉴了较为严谨科学的检索策略[36]。检索式如下:TS = ((“Big Data” or Bigdata) OR (((Big Near/1 Data or Huge Near/1 Data) OR “Massive Data” OR “Huge Information” OR “Big Information” OR “Large -scale Data” OR “Semi -Structured Data” OR “Unstructured Data”) AND (“analytic *” OR “analyz *” OR “analys *”))),检索区间选取为 2008 年至 2016 年,最终得到 17 381 篇论文文献。

关键词在一定程度上可以表征文章的主题,但由于其存在未规范的词汇,且存在同义、近义或无实际意义的词汇,所以需要对关键词进行进一步处理。主要步骤如下:①合并作者关键词(Keywords -Author)和扩展关键词(Keywords -Plus)字段,得到关键词 39 394 个;②通过 VantagePoint 模糊匹配模块[37],消除关键词的单复数形式及词形变化,如“networks”和“network”合并为“network”,得到关键词 35 426 个;③建立人工词表,将部分简写关键词与其全称合并,如“HDF”和“Hierarchical Data Format”合并为“Hierarchical Data Format”,最终得到 35 299 个关键词。

考虑到样本量较小时,聚类结果解释性和适用性较差,样本量太大时,计算聚类结果准确性时需要人工标记的工作量太大,故本文选取了发文量在 10 篇以上的 137 位专家(数据集中共有 47 489 位作者)进行聚类分析。在关键词的选择上应该兼具代表性和全面性,词频在 40 以上的 251 个关键词对文章覆盖率达到 71.7%,故而本文选取了前 251 个关键词对专家进行分类。基于构建的 137 位作者和 251 个关键词的矩阵,可以对此作者 - 关键词矩阵进行聚类计算。

重叠 K-means 算法要求设定聚类个数和选择初始中心点,本文结合大数据领域研究综述的分类体系和关键词主成分分析的结果进行选择。通过阅读大数据领域的研究综述及报告,李贺、袁翠敏等[38]将大数据研究分为 3 个方面:①大数据的基本理论研究,包括大数据的起源与发展,基本概念,数据特征及基本构架,现实意义等;②大数据存储与分析技术研究,包括云计算、Hadoop 和 MapReduce 算法及其改进,数据挖掘聚类技术及其他技术;③大数据应用研究,如应用于生物医药的基因测序,社会网络领域的社交信息挖掘等。2014 年工业和信息化部电信研究院发布的《大数据白皮书》[39]中首先探讨了大数据的概念;在技术层面,认为大数据从数据源经过分析挖掘到最终获得价值一般需要经过 5 个主要环节,包括数据准备、数据存储与管理、计算处理、数据分析和知识展现,其中大数据存储、技术和分析技术是关键;在应用方面,指出大数据的应用处于发展初期,应该予以高度的重视和支持。通过阅读大数据领域的行业报告及综述研究[40]笔者发现,本领域比较权威的分类方法是将研究领域分为 3 类,即大数据的基本理论研究、大数据存储与分析技术研究以及大数据应用研究。同时,本文对大数据领域的词频在 40 以上的前 251 个关键词进行主成分分析,得到 12 个类别,分别为:Classification、Lasso、Recommender System、Hadoop、Hadoop(2)、City、Gene、Managers、Mass spectrometry、Risk、Thing 和 Twitter。聚类结果可视化如图 1 所示。其中 Classification 和 Lasso 两类呈现一定相关性,Recommender System、Hadoop、Hadoop(2)、Gene 和 Mass spectrometry 5 类呈现相关性,City、Managers、Thing 和 Twitter 4 类呈现相关性。由此主成分分析的结果可以看到上述的 12 个主成分类别也主要分为 3 类,这与文献综述及行业报告的分类是一致的。

故而,本文综合主成分分析的结果与大数据领域的研究综述的分类方法,将大数据领域具体分为 3 个类别:①基本理论研究,以 Classification, Lasso 等关键词为代表;②存储与分析处理技术,以 Cloud Compute、Hadoop、MapReduce、Recommender System 等关键词为代表;③应用研究,以 Internet of Thing、Smart City、Twitter、Gene、Manager 等关键词为代表。K-means 算法对初始聚类中心十分敏感,聚类结果随不同的初始输入而波动[41]。本文通过 PCA 主题聚类之后对初始聚类中心代表专家进行选择。A. J. Jara 共发表论文 12 篇,其中 11 篇均以大数据的应用研究为主题,故而选取 A. J. Jara 为应用研究类别的代表。同理,选择 S. Fong 为基本研究领域的代表专家,X. Zhang 为存储与分析处理技术的代表专家。

4.2 实证分析结果

通过重叠 K-means 聚类算法和 TFIDF 加权的重叠 K-means 聚类算法识别出的专家专长如表 1 所示(表中只列出了发文量在前 20 位的专家)。表 1 中的聚类结果按照专家属于不同类别的隶属度大小排序, 如发

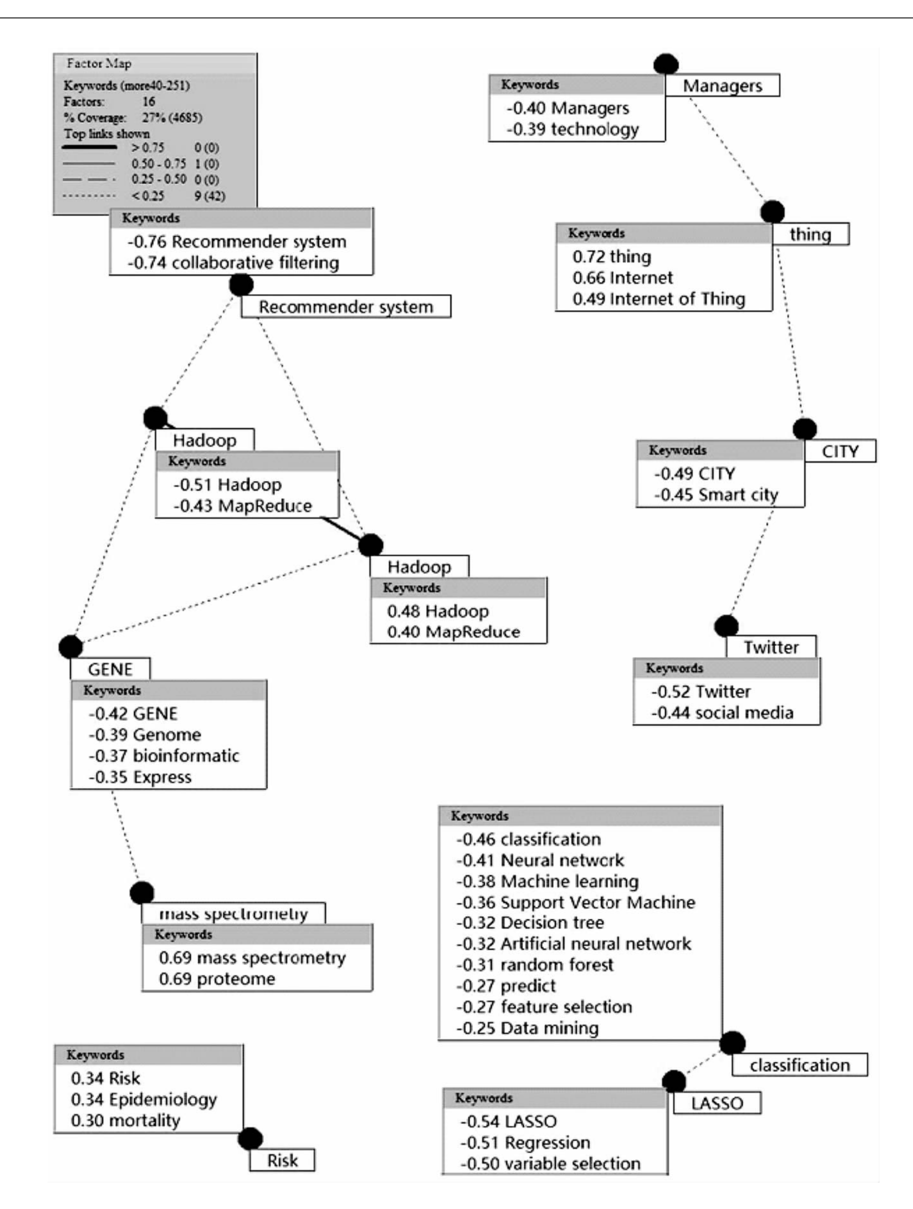


图 1　前 251 个关键词的主成分分析图

文量排名第一的 L. Wang 其聚类结果是 2、1、3，表明其研究专长主要在大数据存储与分析处理技术研究，同时对大数据基本理论研究和应用研究也有所涉猎。从表中可以看出发文量在前两位的专家在三个研究领域都有所涉猎，属于多专长专家。

本文对重叠 K-means 聚类算法和 TFIDF 加权的重叠 K-means 聚类算法得到的聚类结果进行统计分析，统计数据分别如图 2、图 3 所示。本文通过重叠聚类算法共识别出 65 位多专长专家，占全部专家人数的 47.4%，如果使用非重叠聚类进行专家识别，则不能很好地识别出这些多专长专家。比如 S. Yogesh 一方面进行大数据实时存取研究，另一方面也将这些技术应用于智能电网领域，在智能电网领域的大数据存取和处理方面开展了大量研究，如果单单把他归为具有某一研究专长的专家则不符合实际情况。从研究内容上看，发文量较高的这 137 位专家的研究专长主要集中在大数据应用研究，以及大数据存储与分析技术研究。这也与目前大数据相关技术已经从基础研究走入实际应用的现实相一致，智慧城市（smart city）、物联网（Internet of Thing）等新兴技术和产业已经充分应用了大数据的相关研究成果，并不断进行改进。同时从结果中也可以看到，兼具大数据存储与分析技术研究专长与大数据应用研究专长的专家占比达到 33.6%（TFIDF 加权的重叠 K-means 聚类结果中此项占比达到 40.9%），这表示大量的专家在进行技术研究的同时也同样关注把技术转化为实际应用的研究，将目前大数据的存储与处理技术广泛应用于医疗、电子商务、交通、安防、通信等领域和行业。

表 1　专家聚类结果(前 20 位)

专家	人工标记			重叠 K-means 聚类			TFIDF 加权的重叠 K-means 聚类		
L. Wang	2	1	3	2	1	3	2	1	3
R. Ranjan	2	1	3	2	1	3	2	1	3
W. Wang	2	3	–	2	3	–	2	–	–
Cuzzocrea	2	3	–	2	3	–	2	3	–
J. Liu	3	2	–	3	2	–	3	2	–
Y. Zomaya	2	–	–	2	3	1	2	–	–
Liu	2	–	–	1	–	–	1	–	–
L. Wang	3	2	–	3	2	–	3	2	–
G. B. Giannakis	3	1	–	3	–	–	3	–	–
F. Xhafa	2	–	–	2	–	–	2	–	–
X. Zhang	2	–	–	1	–	–	1	–	–
J. Chen	2	–	–	1	–	–	1	–	–
F. Herrera	2	1	–	3	2	1	2	3	1
J. Kepner	2	1	–	2	3	–	2	3	–
X. Cheng	2	–	–	2	–	–	2	–	–
J. Wang	2	–	–	2	3	1	2	–	–
L. T. Yang	2	–	–	2	–	–	2	–	–
W. Dou	2	1	–	1	2	–	2	1	–
V. Gadepally	2	3	–	3	2	–	2	3	–
Y. Liu	3	2	–	3	–	–	3	2	–

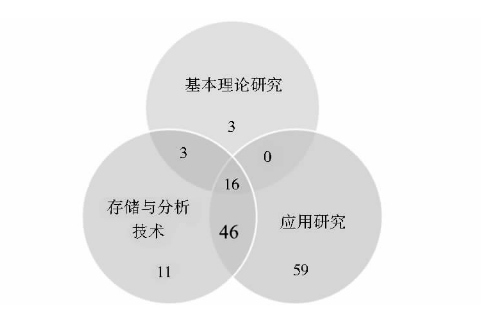


图 2　重叠 K-means 聚类各类别专家数(位)

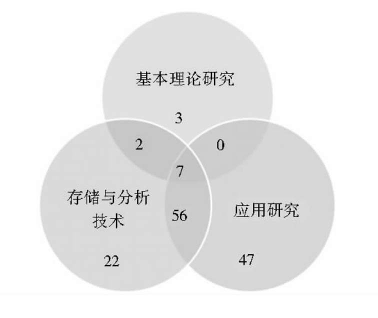


图 3　TFIDF 加权的重叠 K-means 聚类各类别专家数(位)

4.3　实证结果评测

本研究邀请了五位大数据领域的专家对上述的 137 位专家的研究专长进行人工标记,以此作为评价实验结果的标准。首先,笔者向领域专家介绍了本文采用的大数据领域研究的分类标准及依据,明确了三个研究领域的内涵、概念及边界;然后,邀请领域专家对每位作者发表的论文进行研究领域划分,每位作者发表论文的研究领域标记为作者的研究领域。如果五位领域专家对于作者的研究领域标记结果相同,则将此标记结果确定为该作者的研究领域人工标记结果;如果领域专家对于作者的研究领域有异议,即研究领域标记结果不同,则笔者和领域专家将再次阅读有异议的文献,然后通过讨论的方式最终确定该作者研究领域的人工标记结果。

通过以上的方法得到了上述 137 位作者研究领域的人工标记结果,表 1 中列举的人工标记按照专家所发表的不同类别文章数量排序。部分专家在两种聚类方法中得到的聚类结果相同,如发文量排名第一的 L. Wang 发表论文共计 35 篇,涉及基础理论研究,如算法改进;存储与分析处理技术研究,如 G-Hadoop;应用研究,如物联网。同时,两种聚类结果也有一定的差异性,如发文量排名 16 位的 J. Wang 在重叠 K-means 算法中被分到基础理论研究和存储与分析处理技术研究两个类别当中,在 TFIDF 加权的重叠 K-means 算法中被分到存储与分析处理技术研究中,这一分类结果也和人工标记结果相一致。这是因为 TFIDF 表征了一个词的类别区分能力,对关键词进行加权之后,类别区分能力强的词在聚类时有更大的影响,可以优化聚类效

果,避免了重叠K-means算法过多地估计了多专长专家的问题。

本文将两种聚类方法得到的聚类结果与人工标记结果进行比较,计算召回率、查准率和F值(见表2),以此评价聚类结果。

表2 聚类结果评价

组别	指标	人工标记	重叠 K-means	TFIDF 加权的重叠 K-means
1	人数	13	21	12
	召回率	–	69.23%	76.92%
	查准率	–	42.86%	83.33%
	F值	–	52.94%	80.00%
2	人数	104	75	87
	召回率	–	63.46%	73.08%
	查准率	–	88.00%	88.37%
	F值	–	73.74%	80.00%
3	人数	91	121	110
	召回率	–	97.80%	92.31%
	查准率	–	73.55%	77.06%
	F值	–	83.96%	84.00%

从表中可以看到虽然使用重叠K-means算法得到的聚类结果并不尽如人意,但是改进后的TFIDF加权的重叠K-means算法得到的聚类结果在召回率、查准率和F值上都有较好的表现:平均召回率达到81.73%,平均准确率达到83.11%,平均F值达到81.75%。结果说明本文提出的TFIDF加权的重叠K-means算法可以较准确、高效地识别专家专长。

对比两种方法,使用TFIDF加权的重叠K-means算法得到的聚类结果相比于重叠K-means算法有更优异的表现:三组查准率分别提高40.48%、0.37%、3.51%,平均提高14.79%;1、2两组的召回率分别提高7.69%、9.62%,虽然3组的召回率下降了5.49%,但是三组平均而言提高了3.94%;三组F值分别提高27.06%、6.26%、0.04%,平均提高11.12%。专家专长识别的完整性和准确性都有了显著的提升。特别是1组,在使用重叠K-means算法时得到的召回率、查准率和F值均低于70%,这可能与从事大数据基本理论研究的专家数较少,专家之间的区分度低有关。而当对算法进行改进后,1组的查准率由42.86%提高到83.33%,召回率由69.23%提高到76.92%,这一结果表明通过TFIDF对关键词进行加权,增强了其类别区分能力,大大地改善了数据量较少时识别效果不好的问题。

5 结论与展望

作为科学研究的主体,专家在科学研究过程中往往具有多个研究兴趣,他们在从事融合课题或交叉研究时具有不可替代的优势。以往的专家分类方法往往将专家唯一地划分到某一领域,而忽略了多专长专家的识别。为了识别多专长专家,本文运用重叠K-means算法对专家进行聚类划分。为了增强特征在聚类时的区分作用,本文创新性地提出TFIDF加权的重叠K-means算法对专家进行重叠聚类分析。以大数据领域为案例,将SCI/SSCI发文量在10篇以上的137位专家进行重叠聚类,结果表明大多数专家涉猎多个研究方向,其中从事存储与分析处理技术的专家与从事应用研究的专家有较大重叠。TFIDF加权的重叠K-means算法得到的专家聚类划分在查准率、召回率和F值上有很好的表现,可以准确、高效地识别出专家的专长。本研究提出的多专长专家识别方法,弥补了传统专家专长识别研究的不足,实验结果表明这种方法是行之有效的,对于多专长专家的识别有很好的效果。

同时,本研究也存在一些不足。K-means算法本身要求设定聚类数目和初始聚类中心,重叠K-means算法并不能克服这一点。本文依据大数据领域内的研究报告和文献综述对研究专长进行定义并确定聚类数目,但是在分类上粒度较粗,并未深入到具体技术细节,从而识别出的专家专长较为宽泛。在之后的研究中,笔者会对细粒度分类标准下的专家专长识别效果进行研究。同时对初始聚类中心代表进行人工选择,选择不同的初始聚类中心代表是否会对聚类结果产生影响是值得探究的问题。本文中计算专家间距离使用了欧氏距离,在之后的研究中将会对距离的选择进行分析,如余弦距离等是否能更好的表征和刻画专家之间的距离,选择更合适的距离度量方法进行聚类计算,优化聚类结果。

参考文献:

[1] 龙昕. 面向专家检索的社区挖掘研究[D]. 昆明:云南大学,2010.
[2] BEDARD J. Expertise and its relation to audit decision quality [J]. Contemporary accounting research, 2010, 8(1): 198－222.
[3] GLASER R. The nature of expertise. Occasional paper No. 107. [EB/OL]. [2017－05－20]. https://eric. ed. gov/? id = ED261190
[4] YIMAMSEID D, KOBSA A. Expert-finding systems for organizations: problem and domain analysis and the DEMOIR approach [J]. Journal of organizational computing &electronic commerce,

2003, 13(1): 1 –24.

[5] 陆伟,刘杰,秦喜艳. 基于专长词表的图情领域专家检索与评价[J]. 中国图书馆学报, 2010, 36(2): 70 –76.

[6] CLEUZIOU G. An extended version of the k-means method for overlapping clustering[C]// IEEE. Proceedings of 19th international conference on pattern recognition. Tampa: IEEE Press, 2008: 563 –566.

[7] 张晓娟,陆伟,程齐凯. PLSA 在图情领域专家专长识别中的应用[J]. 现代图书情报技术, 2012, 28(2): 76 –81.

[8] APPIO F P, CESARONI F, MININ A D. Visualizing the structure and bridges of the intellectual property management and strategy literature: a document co-citation analysis [J]. Scientometrics, 2014, 101(1): 623 –661.

[9] SONG X, TSENG B L, LIN C Y, et al. ExpertiseNet: relational and evolutionary expert modeling[C]// ARDISSONO L, BRNA P, MITROVIC A. Proceedings of user modeling 2005. Edinburgh: Springer. 2005:99 –108.

[10] YANG K W, HUH S Y. Automatic expert identification using a text categorization technique in knowledge management systems [J]. Expert systems with applications, 2008, 34(2): 1445 –1455.

[11] DOM B, EIRON I, COZZI A, et al. Graph-based ranking algorithms for e-mail expertise analysis[C]//KIM W, KOHAVI R, GEHRKE J, et al. Proceedings of the 8th ACM SIGMOD workshop on research issues in data mining and knowledge discovery. San Diego: IEEE Press,2003:42 –48.

[12] 毛进,李纲. 一种基于 OKM 的研究领域专家图谱构建方法[J]. 图书情报工作, 2014, 58(14): 34 –40.

[13] 魏圆圆,钱平,王儒敬,等. 知识工程中的知识库、本体与专家系统[J]. 计算机系统应用, 2012, 21(10): 220 –223.

[14] 吴春胤,陈壮光,王浩杰,等. 基于本体的专家系统研究综述[J]. 农业网络信息, 2013(4): 5 –8.

[15] 胡月红,刘萍. 基于本体概念的专长表示研究[J]. 图书情报工作, 2012, 56(4): 17 –21.

[16] 刘昕民,桂卫华,杨柳,等. 基于模糊领域本体的专家遴选服务研究[J]. 北京理工大学学报, 2013, 33(5): 484 –489.

[17] LIU R. A new bibliographic coupling measure with descriptive capability[J]. Scientometrics, 2016, 110(2): 1 –21.

[18] LI Y, ZHANG G, FENG Y, et al. An entropy-based social network community detecting method and its application to scientometrics[J]. Scientometrics, 2015, 102(1): 1003 –1017.

[19] 巩军,刘鲁. 基于知识网络的专家知识的表示与度量[J]. 科学学研究, 2010, 28(10): 1521 –1529.

[20] 刘萍,周梦欢. 基于共词网络的专家专长挖掘[J]. 情报科学, 2012, 30(12): 1815 –1819.

[21] STEYVERS M, SMYTH P, ROSENNAZVI M, et al. Probabilistic author-topic models for information discovery[C]//KOHAVI R. Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining. New York:ACM, 2004:306 –315.

[22] TEH Y W, JORDAN M I,BEAL M J, et al. Hierarchical Dirichlet processes[J]. Journal of the American Statistical Association, 2006, 101(476): 1566 –1581.

[23] YAU C K, PORTER A L, NEWMAN N, et al. Clustering scientific documents with topic modeling[J]. Scientometrics, 2014, 100(3): 767 –786.

[24] BLEI D M, JORDAN M, GRIFFITHS T L, et al. Hierarchical topic models and the nested Chinese restaurant process[C]// THRUN S, SAUL KL. Proceedings of the 16th international conference on neural information processing systems. Cambridge: MIT Press, 2003:17 –24.

[25] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of machine learning research, 2003, 3(4/5): 993 –1022.

[26] HOFMANN T. Probabilistic latent semantic indexing[C]// GEY F, HEARST M, TONG R. Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. Berkeley: ACM, 1999: 50 –57.

[27] MULDER W D. Optimal clustering in the context of overlapping cluster analysis[J]. Information sciences, 2013, 223(4): 56 –74.

[28] N'CIR,BEN C E, CLEUZIOU G, et al. Overview of overlapping partitional clustering methods[M]. New York:Springer International Publishing, 2015: 245 –275.

[29] ZHANG Y, PORTER A L, HU Z, et al. "Term clumping" for technical intelligence: a case study on dye-sensitized solar cells [J]. Technological forecasting & social change, 2014, 85(4): 26 –39.

[30] LEE K, JUNG H, SONG M. Subject-method topic network analysis in communication studies[J]. Scientometrics, 2016, 109(3): 1761 –1787.

[31] AHLGREN P, JARNEVING B. Bibliographic coupling, common abstract stems and clustering: a comparison of two document-document similarity approaches in the context of science mapping[J]. Scientometrics, 2008, 76(2): 273 –290.

[32] 刘勘,刘萍. 基于 VSM 的专家领域分析及可视化研究[J]. 图书情报工作, 2011, 55(10): 74 –77.

[33] 施聪莺,徐朝军,杨晓江. TFIDF 算法研究综述[J]. 计算机应用, 2009, 29(6): 167 –170.

[34] 张玉芳,彭时名,吕佳. 基于文本分类 TFIDF 方法的改进与应用[J]. 计算机工程, 2006, 32(19): 76 –78.

[35] N'CIR, BEN C E, ESSOUSSI N. On the extension of K-means for overlapping clustering-average or sum of clusters' representatives? [C]// FRED A, DIETZ J, LIU K. IC3K. KDIR/KMIS 2013 - Proceedings of the international conference on knowledge discovery and information retrievaland the international conference on knowledge management and information sharing. Vilamoura: Springer, 2013:208 –213.

[36] HUANG Y, SCHUEHLE J, PORTER A L, et al. A systematic method to create search strategies for emerging technologies based on the Web of Science: illustrated for 'big data'[J]. Scientometrics, 2015, 105(3): 2005-2022.

[37] VantagePoint[EB/OL]. [2017-05-20]. http://www.thevantagepoint.com/.

[38] 李贺,袁翠敏,李亚峰. 基于文献计量的大数据研究综述[J]. 情报科学, 2014, 32(6): 148-155.

[39] 工业和信息化部电信研究院. 大数据白皮书[R]. 北京:工业和信息化部电信研究院,2014.

[40] CHURYKN T, JANVRIN D, WATSON M. Special issue on big data[J]. Journal of accounting education, 2017, 38(1):1-2.

[41] STEINLEY D. Local optima in K-means clustering: what you don't know may hurt you[J]. Psychological methods, 2003, 8(3): 294-304.

作者贡献说明:

刘晓豫:提出研究思路,设计研究方案,负责采集、清洗和分析数据并撰写文章;

朱东华:参与研究方案的设计;

汪雪锋:参与论文思路确定与论文修订;

黄颖:参与论文修订。

Multi-expertise Researcher Identification: A Case Study of the Big Data

Liu Xiaoyu Zhu Donghua Wang Xuefeng Huang Ying

School of Management and Economics, Beijing Institute of Technology, Beijing 100081

Abstract: [**Purpose/significance**] In response to the rapid shifting of knowledge needs, how to choose the appropriate researchers for a given problem is an important issue for the government, companies, as well as research institutions. When we face a real complex problem, it is essential to find multi-expertise researchers. This research aims to find a proper way to identify multi-expertise researchers. [**Method/process**] This paper used a Term Frequency - Inverse Document Frequency (TFIDF) weighted overlapping K-means clustering method. Based on the researchers' co-authorship network built up from the publication data, the TFIDF weighted overlapping K-means clustering method was applied to cluster researchers into overlapping clusters and identify the multi-expertise researchers. [**Result/conclusion**] Results show that the TFIDF weighted overlapping K-means method has an advantage over the previous work in terms of the precision ratio, the recall ratio and the F-value, so such a method can be beneficial to identify multi-expertise researchers.

Keywords: researcher identification overlapping K-means multi-expertise researcher big data Term Frequency - Inverse Document Frequency (TFIDF)